





Strategic Information Operation in YouTube: The Case of the White Helmets

Nazim Choudhury^(✉) , Kin Wai Ng , and Adriana Iamnitchi 

University of South Florida, Tampa, USA
{nachoudhury, kinwaing, aii}@usf.edu

Abstract. Strategic information operations (e.g. disinformation, political propaganda, and other forms of online manipulation) are critical concerns for researchers in social cyber security. Two strategies, spoofing and astroturfing, are often employed in disinformation campaigns to discredit truthful narratives, confuse audiences, and manipulate opinions by censoring critics' voices or using one-sided testimonials. This study analyzes patterns of spoofing and astroturfing on YouTube regarding narratives about the White Helmets, a Syrian civil volunteer organization active in the long armed conflict in the country.

Keywords: Information operations · Spoofing · Astroturfing · YouTube

1 Introduction

The Syrian civil war transcended from the battlefield to the digital information space seeking to harness the dissemination of old-fashioned propaganda using various digital information media, including YouTube. The objective of the digital campaign was to manipulate the public perception of the civil war and elicit support from foreign, especially western sympathizers. Since September 2016, pro-Assad supporters and their allies mounted a massive disinformation campaign to discredit a civilian rescue group named 'White Helmets' (WH). The group's humanitarian activities, its efforts to document the targeting of civilians through video evidence, and its refusal to align itself with any other group or military factions engaged in the complex Syrian conflict put the group at odds with the government and their allies, including Russia [12].

Political propaganda, disinformation, influence and information manipulation, also known as Strategic Information Operations [16], denote efforts to manipulate public opinions and perceptions of events via intentional alteration of the information environment. Proliferation of social media platforms and freedom of integrating YouTube videos in tweets, blogs, or social media posts along with YouTube's recommendation algorithms have been exploited to frame public discourse and conduct strategic information operations. Wilson and Starbird [18] presented a tracing of information trajectories across YouTube, Twitter, and non-mainstream news domains to demonstrate how state-sponsored

media apparatus shaped the disinformation campaign against the White Helmets. Two conventional techniques for the manipulation of information environments are spoofing and astroturfing. Spoofing involves deception and trickery to misrepresent both source identity and information veracity through falsification, suppression, or amplification [5]. Astroturfing, on the other hand, involves manufactured, deceptive and strategic top-down activities initiated by politically or ideologically-motivated actors to mimic bottom-up activity [11]. This can involve autonomous or non-autonomous individuals who create an illusion of widespread support for a candidate or opinion. Spoofing was found on Twitter [4], while astroturfing was extensively found on Twitter [8,9], Facebook [6], emails [10] and websites [19]. Platform-specific features (such as type of content and type of user interactions) suggest different ways of recognizing astroturfing and spoofing. For example, in Twitter, user attributes and their temporal activities were used to identify spoofed accounts [17], whereas their interactions patterns (e.g., co-tweet and co-retweet) were used to identify astroturfing [9]. Hussain et al. [3] analyzed user engagement on YouTube and applied social network analysis techniques to identify inorganic behaviors.

This article conducts interpretative analyses on four months of YouTube data focusing on the strategic information operations mounted against the WH. Our contributions are twofold: first, we show evidence of inorganic behaviors in user comments on WH-related videos which suggests the possibility of information and identity spoofing. We also highlight unusual correlation between highly visible videos and highly appreciated comments by a small group of users and suspect the existence of astroturfing activities. Second, the methods we present, while inspired by previous work, have been adapted to the particularities of YouTube and can serve for similar studies on this platform.

2 Dataset

For this study, we focus on YouTube data specific to the White Helmets, collected using the YouTube API keyword query tool between April 1, 2018 and July 31, 2018. The dataset was provided privately as part of DARPA SocialSim program. The list of keywords used in the YouTube data collection are ‘white helmets’, ‘cascos blancos’, ‘capacetes brancos’, ‘caschi bianchi’, ‘casques blancs’, ‘elmetti bianchi’, ‘weisshelme’, ‘weiß helme’, ‘syrian civil defence’, and lastly White Helmets in Russian and Arabic. This dataset contains 76 videos from 42 different channels. For each video, we have their corresponding top-comments and replies. Overall, the dataset includes a total of 10,636 comments and 3,633 replies done by 8,071 and 1,892 users, respectively. The total counts of likes (liked by the other commenters) associated with each comment or reply were also collected.

3 Spoofing

As a technique of disinformation, Martin [5] identified two master types of spoofing: identity and information. Identity spoofing involves constructing false digital

identities. Information spoofing involves misrepresenting the content of a message through processes of falsification, suppression, amplification or even sometimes blending these techniques together. In order to understand the extent to which identity and information spoofing contributed to the anti-WH campaign on YouTube, we considered the observations in [2] on analyzing the Twitter activities of specialized bot accounts, known as social bots, during the Maidan protest in Ukraine. The authors found that these social bots not only were capable of mimicking humans and thus hide detection via constantly changing the contents of their tweets, but they also promoted certain topics through massive repetition of messages and retweeting of selected tweets.

Instead of focusing on the behavioral characteristics of users accounts, following Schäfer et al. [14], we used a corpus-linguistic approach to detect bots among YouTube commenters. The collected YouTube comments on different videos uploaded by different channels were normalized via cleaning (removing white spaces, punctuation, URLs and emojis), and tokenization. A hash structure via locality sensitive hashing (LSH) [15] with minHash [13] was used to recognize semantic and near-duplicates of the normalized comments. LSH aims to preserve the local relations of the data by significantly reducing the dimensionality of the dataset and ensuring hash collisions for similar items, something that hashing algorithms usually try to avoid. Through the hash structure, we identified near-duplicated comments posted by a small number of users on different videos. In general, near-duplicate comments contained similar texts with small modifications predominantly generated via adding/changing emojis, URLs, or word(s) in the front or end of the comments.

Figure 1 presents network snapshots of videos-comments and videos-commenters constructed by considering the near-duplicate comments posted on different videos. In the left network, the red nodes are videos and the blue nodes denote *top comments* (top-ranked popular comments visible on the first page of all comments) on those videos. The green edges connect a video with its comments and the orange edges represent the near-duplicate relationship between comments. On the right, we present a snapshot of the video-commenters relationships where the purple nodes are the commenters responsible for the near-duplicate comments and the red nodes represent the videos. The thickness of the orange edges is proportional to the number of posts (comments or replies to other users' comments) made by that user on that particular video. Thus, the network on the left presents videos as they are (indirectly) connected by duplicate posts and the network on the right presents the very same videos connected by the user accounts who post the duplicate messages.

Similar to the findings reported in [3], we find inorganic commenting behaviors: First, out of the 76 videos in the initial dataset, 62 appear in Fig. 1 (top) connected by a total of 241 posts with at least one duplicate (out of a total of more than 14K comments and replies), of which only 93 are distinct according to our duplicate evaluations. These messages are posted by 87 users. The most common duplicate message is posted 19 times by the same user on the same video, as depicted by the highly connected cluster in the left network. There are 76 messages that are posted twice in this small dataset. Second, the same

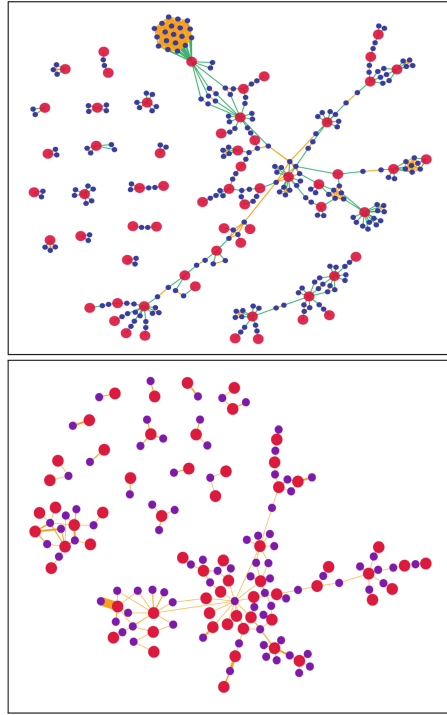


Fig. 1. Videos-comments network (top): red nodes represent videos and blue nodes represent comments; green edges denote a comment on a video and orange edges connect duplicated comments. The corresponding videos-commenters network (bottom): purple nodes represent the user accounts who made the comments in the left network snapshot. The weight of the orange edges is the number of posts made by a user on a video. (Color figure online)

comment is posted repeatedly by the same user on the same video. For example, the 19 duplicate messages posted on the same video are posted by the same user. In another example, one user engages with a video 99 times via comments and replies, as shown by the thick edge in the network on the right. And third, duplicate messages are posted by different user accounts. We found five videos with evidence of identity spoofing. Each such video has pairs of duplicated comments in which duplicate messages are posted by different user accounts. Three of these videos have one such pair of duplicated messages, the fourth has two pairs, and the fifth has three pairs of duplicated messages allegedly authored by different user accounts. Because of the inorganic behavior of the users involved in posting duplicate messages, we will use ‘spoofed accounts’ and ‘social bots’ interchangeably to refer to these identified commenters in the rest of the study. A deeper investigation in the behavioral patterns around identity and information spoofing led us to discover the following three types of activity:

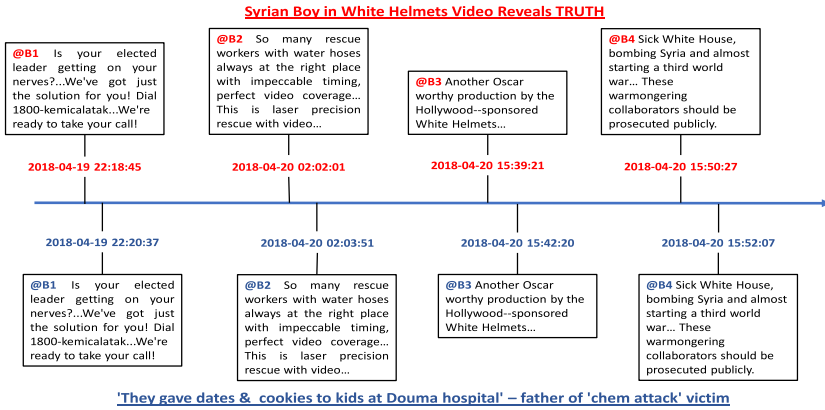


Fig. 2. Timeline of coordinated commenting behaviors by four spoofed accounts (B1, B2, B3, B4) on two videos (titled in red and blue colored texts) shared by the channels named ‘Lift the Veil’ and ‘RT’, respectively. (Color figure online)

Coordination: We identified the coordination of multiple spoofed accounts who posted comments on two YouTube videos. Figure 2 presents the timeline of comments by four user accounts (shown as B1, B2, B3, B4) on two videos uploaded on two different channels. The first video was hosted by RT, a Russian state-sponsored media channel, and the second one by ‘Lift the Veil’, a channel mostly dedicated to sharing conspiracy theories. Each commenter, identified in our analysis as spoofed accounts, posted the same message on the two videos at approximately 2-min intervals. Even more unusual is that in both YouTube videos, with different titles on the same topic, these users commented in the same order and with similar inter-arrival time between their postings. This suggests the possibility of a coordinated effort in which spoofed accounts controlled by the same agent promote messages on different videos posted on different channels with the common objective of discrediting WH.

DogPiling: Figure 3 (top) shows two users (green nodes) who posted comments on two videos and who received many reactions (as replies to those comments) from other users. The two videos are hosted by RT and Vesti News channels—a brand used by the Russian broadcaster VGTRK (All-Russia State Television and Radio Broadcasting Company). The arrow directions represent comments to videos or replies to comments. Comments by the green nodes, denouncing/criticizing the Syrian regime and their collaborators, were explicitly challenged and confronted by multiple replies from both spoofed and non-spoofed accounts in a cascade of dissent and insults. This form of attack is a form of “dogpiling” [7] which manifests via hostile efforts to drown out and drive out (through intimidation and harassment) users and content who, in this instance, showed support to the WH.

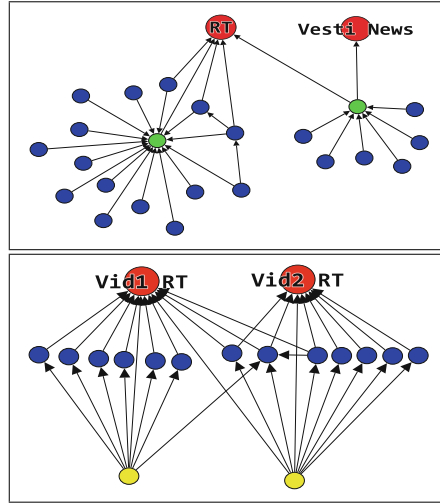


Fig. 3. Dogpiling network (top): Dissenting comments on a video (green nodes) are confronted by multiple users. Replicators network (bottom): Social bots (yellow nodes) replicate and modify the comments by replacing keywords. (Color figure online)

Replication: The third type of attack is portrayed in Fig. 3 (bottom), where two social bots were found replying to multiple commenters. While the commenters were discrediting the Assad regime and its allies as well as the corresponding YouTube channel that hosted the videos, the repliers replicated the same message only by replacing keywords. For example, the comment ‘everyone knows Russia is a state of liars’ was replied with the text ‘everyone knows USA is a state of liars’. The average time spent on bulk replies against 10–12 posts was 4 min in two instances. This may suggest automated behavior or a dedicated human user whose comments are triggered by this type of messaging.

4 Astroturfing?

Online astroturfing creates the illusion of widespread grassroots support by posting or artificially supporting messages that advance a specific opinion or narrative [20]. Astroturfing is a common strategy of disinformation in politics [9] where the goal is to create the impression of popular support.

Due to the anonymity offered by social media platforms, astroturfing is usually covert, sophisticated, and hard to distinguish from genuine grassroots support. To uncover online astroturfing in the campaign against WH, we concentrated on the top commenters with high volume of comments on different YouTube channels. We represent a $n \times r$ matrix with $n = 200$ rows for each commenter who posted comments on different videos uploaded by $r = 27$ channels between April 1st 2018 and July 31st 2018. Figure 4 (top) visualizes this matrix: the red dots represent comments posted by each of the 200 users on

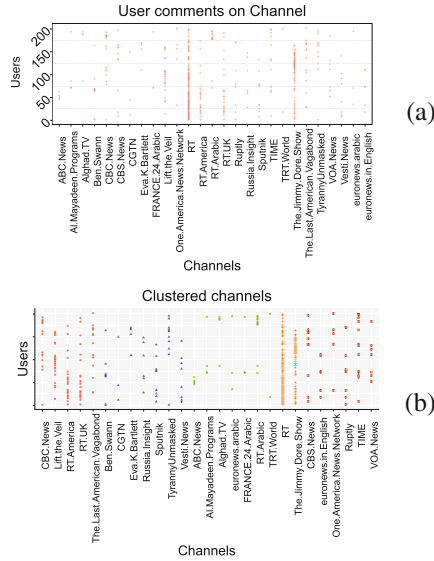


Fig. 4. (a) Commenting patterns by the top-200 active commenters on different YouTube channels (top) and (b) clusters of channels obtained by block clustering. Color codes represent clusters of channels (bottom).

Table 1. Popularity achieved by the videos published in each cluster, identified by the block-clustering approach. Popularity metrics are presented as the average likes of the comments and average likes, and views of the videos.

Cluster	Videos	Spoofed accounts	Views	Likes
C1	11	5	175,971	5,750
C2	8	3	103,197	4,405
C3	10	5	210,819	1,534
C4	10	11	660,383	21,649
C5	9	3	34,208	746

each channel. An entry in m_{xy} is equal to one if commenter x commented on a video uploaded by channel y and zero otherwise. An expectation-maximization algorithm based on a block mixture model [1] was used to cluster the channels by user interactions. Figure 4 (bottom) presents the optimal clustering of the channels according to user interactions. We obtained five such clusters. Clusters C1, C3, and C5 (as per Table 1) contain a mixture of main stream media and questionable sources (conspiracy theories channels or subsidiaries of Russia Today). Thus, C1 contains *CBC News* and four questionable channels: two known for promoting conspiracy theories, *Lift the Veil* and *The Last American Vagabond*, and two subsidiaries of Russia Today, *RT America* and *RT UK*. C3

contains *ABC News*, *Al Mayadeen Programs*, *Alghad TV*, *FRANCE 24 Arabic*, *RT Arabic*, *TRT World*, *euronews arabic*. C5 contains *CBS News*, *One America News Network*, *Ruptly*, *TIME*, *VOA News*, and *euronews in English*. The other two clusters obtained are more homogeneous, containing channels of well-known anti-WH activists along with news channels classified as heavily biased. C2 contains the YouTube channels *Ben Swann*, *CGTN*, *Eva K Bartlett*, *Russia Insight*, *Sputnik*, *TyrannyUnmasked*, *Vesti News*. C4 contains *RT* and *The Jimmy Dore Show*.

As shown in Table 1, C4 (that consists of the Russia Today channel and a far-left YouTube channel run by the American stand-up comedian and political commentator Jimmy Dore) is the most popular cluster: it attracted the largest number of manipulative accounts as detected by our analysis from the top 200 commenters; and it generated the largest number of likes and views on average per video. Most messages refer to the video titled “Carla Ortiz Shocking Video From Syria Contradicts Corp News Coverage”, where Jimmy Dore interviewed Carla Ortiz, a former Hollywood actress and a front-runner anti-WH activist. We will refer to this video as the “Interview” in the rest of the paper.

The “Interview” was posted 2 days after the last video in the RT channel was posted. 110 users commented on videos published by both channels. From these, 60 users commented on RT’s videos before they commented on the Jimmy Dore Show’s video. We will focus on these 60 users and we will refer to them as *RT-commenters*, as they appear to have engaged earlier with the anti-WH narratives as promoted by the Russian government-sponsored RT channel. 34 users commented on the “Interview” video for the first time in our dataset. The “Interview” video had the highest number of views (205,814) and likes (10,437) in cluster C4, followed by one of the videos from RT with much less popularity: 110,258 views and 3,005 likes. To put things in context, the number of subscribers to the Jimmy Dore’s show channel is seven times smaller (567,444 subscribers) compared to the RT channel (3,560,843 subscribers), yet its one video in our dataset received much more attention. Surprisingly, in terms of average number of likes to comments made on the “Interview”, the users who commented in both channels received more appreciation (18.8 likes per comment on average) than the users who only commented on the “Interview” (9.19 likes per comment on average). Moreover, the 60 RT-commenters had an even higher average of 24.38 likes per comment for their comments on the “Interview”. What makes these dedicated commenters of RT-posted videos be so appreciated on the Jimmy Dore Show’s anti-WH video, especially in comparison with the other commenters who never commented on the RT-channel videos?

To better understand the appeal of these RT-commenters, we look at how their comments are appreciated on other channels outside the C4 cluster. In cluster C1 where we observed four questionable channels along with Canadian Broadcasting Corporation (CBC), 34 RT-commenters commented on all 11 videos and achieved an average 2.82 likes against an average of 1.45 likes acquired by the other commenters. Similarly, in cluster C2, where anti-WH self-proclaimed journalists have their own YouTube channels, 16 RT-commenters commented on all

eight videos and achieved 8.62 likes on average per comment as opposed to 1.49 likes acquired by the non-RT commenters. In cluster C3, however, we observe a contrasting phenomenon: only one RT-commenter posted a single comment. Most of the channels in this cluster are Assad regime-supporting Arabic language news channels. In cluster C5, we identified 12 RT-commenters who commented on three videos published by the CBS, Euronews in English, and the Time magazine’s YouTube channels, respectively. The video published by CBS presented the news of halting US funds to the White Helmets. This was also a chosen narrative by the anti-WH community to defund the humanitarian group. The video published by the Euronews in English and Time had title *Israel Is Evacuating Hundreds of White Helmets Rescue Workers from Syria*, where the Israeli military in coordination with its US and European allies evacuated hundreds of Syrian rescue workers known as the White Helmets from near its volatile frontier with Syria, in a complex and first-of-a-kind operation. Cooperation with the Israeli army and the West is highly denounced in the Middle Eastern political context and was used against WH to portray the group as anti-Muslim. These three videos achieved the highest view counts in that cluster (CBS 6,941, Euronews 4,636, and Time 4,607).

We thus observe that the videos on which the RT-commenters comment have a large number of views independent of the channel in which they are posted (whether mainstream media or conspiracy theory promoters). At the same time, we observe that wherever they comment, the RT-commenters receive on average many more likes than the rest of the commenters in that channel. One possible explanation may be the existence of astroturfing: a concerted effort of promoting some videos by manipulating the YouTube algorithms via comments and likes to comments. There are other possible explanations of this correlation between the high popularity of this group of users and the popularity of the videos with which they interact. We do not claim to distinguish between correlation and causality in this work, thus the question mark in the title of this section. In this work we only highlight this unusual correlation and plan to investigate the causality question in future work.

5 Conclusions

We analyzed four months of YouTube data in an attempt to identify strategic information operations against the White Helmets. We used a corpus linguistic approach and hash structure to uncover patterns of identity and information spoofing. We discovered many instances of the same message (with slight variations in emojis or other embellishments) being repeatedly posted on the same video or on different videos. We also discovered instances where different user accounts post the same message on the same or different videos. We also found three patterns of antagonistic behavior demonstrated by spoofed accounts: coordinated attacks, dogpiling, and automated message replication.

In addition, we used a block clustering approach to cluster the channels that upload WH-related videos based on the user engagement. In the absence of

ground truth data, we used three performance metrics (i.e., numbers of video likes, video views and comment likes) to measure the popularity of the videos uploaded to different channels and their associated comments. We found that a small group of users engaged heavily with videos published by channels that support anti-WH narratives promoted by RT channels. Not only that this small group has received higher attention to their comments and replies, but also the popularity of these videos is significantly higher than the videos where these users did not engage. We suspect this unusual behavior can be the effect of astroturfing, in which a small number of user accounts coordinate to exploit the YouTube video promotion algorithm to reach a wider audience and promote specific narratives. We will test this hypothesis in future work.

Acknowledgements. This work is supported by the DARPA SocialSim Program and the Air Force Research Laboratory under contract FA8650-18-C-7825. The authors would like to thank Leidos for providing data.

References

1. Govaert, G., Nadif, M.: An EM algorithm for the block mixture model. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(4), 643–647 (2005)
2. Hegelich, S., Janetzko, D.: Are social bots on Twitter political actors? Empirical evidence from a Ukrainian social botnet. In: Tenth International AAAI Conference on Web and Social Media (2016)
3. Hussain, M.N., Tokdemir, S., Agarwal, N., Al-Khateeb, S.: Analyzing disinformation and crowd manipulation tactics on YouTube. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1092–1095. IEEE (2018)
4. Innes, M.: Techniques of disinformation: constructing and communicating “soft facts” after terrorism. *British J. Sociol.* **71**, 284–299 (2020)
5. Innes, M., Dobрева, D., Innes, H.: Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing. *Contemp. Soc. Sci.* 1–15 (2019)
6. Isaac, M.: Facebook finds new disinformation campaigns and braces for 2020 torrent, October 2019. <https://www.nytimes.com/2019/10/21/technology/facebook-disinformation-russia-iran.html>
7. Jhaver, S., Ghoshal, S., Bruckman, A., Gilbert, E.: Online harassment and content moderation: the case of blocklists. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **25**(2), 1–33 (2018)
8. Keller, F.B., Schoch, D., Stier, S., Yang, J.: How to manipulate social media: analyzing political astroturfing using ground truth data from South Korea. In: Eleventh International AAAI Conference on Web and Social Media (2017)
9. Keller, F.B., Schoch, D., Stier, S., Yang, J.: Political astroturfing on Twitter: how to coordinate a disinformation campaign. *Polit. Commun.* **37**(2), 256–280 (2019)
10. King, G., Pan, J., Roberts, M.E.: How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Polit. Sci. Rev.* **111**(3), 484–501 (2017)
11. Kovic, M., Rauchfleisch, A., Sele, M., Caspar, C.: Digital astroturfing in politics: definition, typology, and countermeasures. *Stud. Commun. Sci.* **18**(1), 69–85 (2018)

12. Levinger, M.: Master narratives of disinformation campaigns. *J. Int. Aff.* **71**(1.5), 125–134 (2018)
13. Rajaraman, A., Ullman, J.D.: *Mining of Massive Datasets*. Cambridge University Press, Cambridge (2011)
14. Schäfer, F., Evert, S., Heinrich, P.: Japan's 2014 general election: political bots, right-wing internet activism, and prime minister shinzō abe's hidden nationalist agenda. *Big Data* **5**(4), 294–309 (2017)
15. Slaney, M., Casey, M.: Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *IEEE Signal Process. Mag.* **25**(2), 128–131 (2008)
16. Starbird, K., Arif, A., Wilson, T.: Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. *Proc. ACM Hum. Comput. Interact.* **3**(CSCW), 127 (2019)
17. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: *Eleventh International AAAI Conference on Web and Social Media* (2017)
18. Wilson, T., Starbird, K.: Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review* **1**(1) (2020)
19. Zerback, T., Töpfl, F., Knöpfle, M.: The disconcerting potential of online disinformation: persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media Soc.* (2020). <https://doi.org/10.1177/1461444820908530>
20. Zhang, J., Carpenter, D., Ko, M.: Online astroturfing: a theoretical perspective (2013)